REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Nov 14, 2000	Final Report, 6/15/99 — 6/15/00

4. TITLE AND SUBTITLE
A Perception and Nonlinear PDE Based Approach to Processing Spoken Words

5. FUNDING NUMBERS
DAAD 19-99-1-0248

6. AUTHOR(S)
Jack Xin and Yingyong Qi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Arizona
University of Arizona, Sponsored Projects Services, 888 N. Euclid Ave #510, Tucson, AZ 85722.

8. PERFORMING ORGANIZATION REPORT NUMBER
FRS 314900

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO Number: 40142.2-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT	12 b. DISTRIBUTION CODE
Approved for public release; distribution unlimited.	

13. ABSTRACT (Maximum 200 words)

please see the attached.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Perception, Nonlinear PDE, Nonlinear Processing, Spoken Words	5
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

20010117 093

A Perception and Nonlinear PDE Based Approach to Processing Spoken Words

Final Report on

Grant DAAD 19-99-1-0248

PI: Jack Xin, Co-PI: Yingyong Qi

Abstract

During the period of June 1999 to June 2000, supported by ARO grant DAAD 19-99-1-0248, we developed a novel nonlinear transformation to process spoken words in noisy environment, based on human hearing perception and properties of focusing partial differential equation (PDE). The transformation was made on the short-term Fourier spectra of speech signals. It was designed to reduce noise through time adaptation, and enhance spectral peaks (formants) by evolving a focusing quadratic Cahn-Hillard equation. Time adaptation and peak focusing (a.k.a lateral inhibition) are essential processing mechanisms in human cochleas.

Numerical results on noisy spoken words indicated that the transformed spectral pattern of the spoken words was insensitive to noise (signal-to-noise ratio (SNR) ranging from 0 to 20 dB). The spectral distances between noisy and original words decreased after the transformation. Numerical experiment on eleven spoken words at SNR = 5 dB, for example, reached a recognition rate as high as 100%. These very encouraging results showed the success of our nonlinear transformation and the needs of its further development within our framework.

In this final report, we state the problem studied, summarize main results, and point out future directions.

1 Statement of the Problem Studied

Consider the spectral data arising from short-time window Fourier transform of a sound wave of a spoken word. The spectral data, known as spectrogram, is a matrix $a_{i,j} = a(t_i, x_j)$, $1 \leq i \leq I$, $1 \leq j \leq J$, where I, J are positive integers; t_i's are discrete sampling of time, typically at a rate of 10-30 ms per frame; and x_j's are discrete sampling of the frequency, typically scaled according to human auditory perception or bark-scale (roughly logarithmic scale). The spectrogram is a space-time distribution of acoustic energy, and as usual, we consider it in unit of decibel (dB), or the $u_{i,j} = 10(\log_{10} |a_{i,j}|^2 - \log_{10} I_0)$, where I_0 is a reference intensity.

Our processing and the resulting recognition will be on the matrix u_{ij} for a spoken word from a vocabulary of words under noisy conditions. *The studied problem is: If u_{ij} is received from a noisy environment, how do we process it so that the essential speech features are enhanced, noise effects are reduced, and recognition rates are improved ?*

It is known that using noisy spectrogram without processing will lead to drastic increase of errors. Methods for reducing noise effects based on auditory perception have been proposed and implemented in the engineering community [4], [11], [12] among others. These methods tend to rely on various ad hoc procedures to model two mechanisms in human audition: (1) time adaptation (reducing redundancy of slowly varying spectrum at any fixed frequency), and (2) spectral peak (formants) isolation, tracking, and enhancement.

A major portion of our study is to develop a systematic and efficient mathematical method to *capture both time adaptation and formant structures* in a noisy speech spectrum without performing detailed case dependent filters or structure tracking.

2 Summary of Most Important Results

Adaptation basically is to reduce any portion of spectrogram at a fixed frequency if there is not enough variation in time, a process occuring in human hearing to remove redundancy. For clean signal, the spectral curve is smooth in time, and so one can use derivative to measure this variation. Due to noise, the spectrogram are often rough, and one must devise an alternative measure of variation. We first divide the frequencies into three (low, mid and high) bands and define an average curve for each band as a representative. Such a three band division is based on human hearing response to multi-frequency stimuli (critical bandwidth) [10]. We then construct an upper and a lower envelope for each band representative so that the difference of the two envelopes is a good measure of true signal variation in time for each band. When

the difference is below a threshold (2dB), indicating either noise or redundant signal, adaptation takes place. This is the nonlinear transformation in time, call it A_d.

For each fixed time, spectral peaks appear in vowels at isolated frequency locations and with decreasing magnitudes towards higher frequencies. These spectral peaks are called formants, and are locations of energy concentration. Enhancing these peaks at the expense of reducing energies at neighboring points is analogous to the well-known lateral inhibition phenomenon in psychoacoustics [5] and [6]. For two tones, one stronger than the other, Houtgast ([5], [6]) showed experimentally that the stronger tone suppresses the nearby weak tone, and attributed this to processes within the cochlear and to the nonlinear aspect of neural coding of sound spectrum. Neural projection of sound spectrum tends to sharpen formant peaks which helps vowel discrimination and recognition. Motivated by this connection, we found and implemented a nonlinear transformation based on evolving a focusing fourth order nonlinear partial differential equation (known as the Cahn-Hillard equation [2], [3], [1]):

$$u_\tau = -\alpha(u^2)_{xx} - \epsilon u_{xxxx}, \tag{1}$$

where $\alpha \geq 1$, $\epsilon > 0$ is small enough. Here τ is the processing time, x is the frequency. To handle roughness of spectral curve resulting from noise effects (or to minimize noise induced spurious peaks), we first take a logarithm on the input (or initial data of (1)), evolve it under (1) up to a time $t = T$, then exponentiate the solution. The nonlinear mapping, call it F_o, is then the composition of these three steps. The focusing step resulting from evolving (1) mimics lateral inhibition effects. Fixing the parameters α, ϵ, and T requires training on clean data set. The transformation F_o is performed at times during a vowel, and hence processes spectrogram along frequency axis. One does not need to know where the peaks are (no tracking is necessary), and the peaks are captured and enhanced automatically by evolving (1). This is the advantage of our method over tracking and filtering method in [11] [12].

The entire nonlinear transformation, call it N, is the combination of time adaptation transformation A_d and the peak focusing transformation F_o. Our numerical experiments on spoken words, with noise added at signal-to-noise ratio (SNR) from 0dB to 5dB, demonstrated that the nonlinear transformation is robust and noise insensitive. Moreover, spectral L^2 distances between noisy words and original words decreased after the transformation. A numerical experiment was performed on eleven spoken words at SNR = 5 dB. A noisy word is recognized numerically by computing the closest L^2 spectral distance from the clean template. The experiment reached a recognition rate as high as 100%. See [9] for analyses and justification on the properties of the transformation, and details of the reported results.

In the future, we plan to investigate lateral inhibition effects by seeking a more neurophysiologically based model ([8],[7]), extending the two tone (or near neighbor) inhibition picture to multitone nonlinear interaction, this is currently under progress.

The other less known but seemingly rather important aspect of method design is to find out how to couple adaptation and lateral inhibition in a nonlinear fashion for better modeling the continuous variation of essential speech spectral information in time. We believe these are fundamental steps towards further improving recognition performance by a better approximation of human auditory processing.

3 List of Publications and Reports

(1) Y-Y Qi, J. Xin, *A Perception and PDE Based Nonlinear Transformation for Processing Spoken Words (current version)*, to appear in Physica D (Nonlinear Phenomena), 2001. See also http://www.ma.utexas.edu/~jxin/speech.html.

(2) Y-Y Qi, J. Xin, *A Perception and PDE Based Nonlinear Transformation for Processing Spoken Words (abridged current version)*, in Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP), Vol. I, pp 445-448, 2000.

(3) ARO Interim Report: *A Perception and PDE Based Nonlinear Transformation for Processing Spoken Words (draft version).*

References

[1] A. Bernoff, A. Bertozzi, *Singularities in a modified Kuramoto-Sivarshinsky equation describing interface motion for phase transition*, Physica D 85(1995), pp 375-404.

[2] J. Cahn, J. Hillard, *Free energy of a nonuniform system,I, Interfacial free energy*, J. Chem. Phys, 28(1958), pp 258-267.

[3] C. Elliot, S. Zheng, *On the Cahn-Hillard Equation*, Arch. Rat. Mech and Analys, 96(1986), pp339-357.

[4] H. Hermansky, N. Morgan, *RASTA Processing of Speech*, IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, 1994, pp 578-589.

[5] T. Houtgast, *Lateral Suppression in Hearing: a psychophysical study on the ear's capability to preserve and enhance spectral contrasts*, research monograph, published by Academische Pers B.V, Amsterdam, 1974.

[6] T. Houtgast, *Auditory Analysis of Vowel-Like Sounds*, Acustica, Vol. 31 (1974), pp 320-324.

[7] B. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 4th edition, 1997.

[8] A. Popper, R. Fay, *The Mammalian Auditory Pathway: Neurophysiology*, Springer Handbook of Auditory Research, eds. R. Fay, A. Popper, Springer-Verlag, 1992.

[9] Y-Y Qi, J. Xin, *A Perception and PDE Based Nonlinear Transformation for Processing Spoken Words (current version)*, to appear in Physica D (Nonlinear Phenomena). See also http://www.ma.utexas.edu/~jxin/speech.html.

[10] L. Rabiner, B-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[11] B. Strope, A. Alwan, *A Model of Dynamic Auditory Perception and Its Application to Robust Word Recognition*, IEEE Transactions on Speech and Audio Processing, vol. 5, no. 5, 1997, pp 451-464.

[12] B. Strope, A. Alwan, *Robust Word Recognition Using Threaded Spectral Peaks*, IEEE Int. Acous. Speech Signal Proc. (ICASSP), Vol. 2, pp 625-628, 1998, Seattle, Washington.